Modeling Multimodal Cues in a Deep Learning-Based Framework for Emotion Recognition in the Wild

Stefano Pini
University of Modena and Reggio Emilia
Modena, Italy
179134@studenti.unimore.it

Olfa Ben Ahmed
EURECOM
Sophia-Antipolis, France
olfa.ben-ahmed@eurecom.fr

Marcella Cornia
University of Modena and Reggio Emilia
Modena, Italy
marcella.cornia@unimore.it

Lorenzo Baraldi
University of Modena and Reggio Emilia
Modena, Italy
lorenzo.balaradi@unimore.it

Rita Cucchiara
University of Modena and Reggio Emilia
Modena, Italy
rita.cucchiara@unimore.it

Benoit Huet
EURECOM
Sophia-Antipolis, France
benoit.huet@eurecom.fr

ABSTRACT

In this paper, we propose a multimodal deep learning architecture for emotion recognition in video regarding our participation to the audio-video based sub-challenge of the Emotion Recognition in the Wild 2017 challenge. Our model combines cues from multiple video modalities, including static facial features, motion patterns related to the evolution of the human expression over time, and audio information. Specifically, it is composed of three sub-networks trained separately: the first and second ones extract static visual features and dynamic patterns through 2D and 3D Convolutional Neural Networks (CNN), while the third one consists in a pretrained audio network which is used to extract useful deep acoustic signals from video. In the audio branch, we also apply Long Short Term Memory (LSTM) networks in order to capture the temporal evolution of the audio features. To identify and exploit possible relationships among different modalities, we propose a fusion network that merges cues from the different modalities in one representation. The proposed architecture outperforms the challenge baselines (38.81% and 40.47%): we achieve an accuracy of 50.39% and 49.92% respectively on the validation and the testing data.

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding*; *Neural networks*;

KEYWORDS

Emotion Recognition, EmotiW 2017 Challenge, Multimodal Deep Learning, Convolutional Neural Networks

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143006

ACM Reference Format:
Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet. 2017. Modeling Multimodal Cues in a Deep Learning-Based Framework for Emotion Recognition in the Wild. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3136755.3143006

1 INTRODUCTION

Emotion recognition is an active research topic in the affective computing community. During the last decade, emotion recognition systems have been integrated in a number of applications across a growing number of domain fields such as cognitive science [31], clinical diagnosis [15], entertainment [38] and human-machine interaction [3]. Automatic emotion analysis and recognition in real-world videos (*i.e.* in the wild) is nevertheless still an open challenge in computer vision. One fundamental limiting factor is that there is almost no large dataset with real-world facial expressions available for emotion recognition. Other challenging factors include head pose variation, complex facial expression variations, different illumination conditions and face occlusion.

Recent achievements in the field are based on the use of data coming from multiple modalities, such as facial and vocal expressions. Indeed, each modality presents very distinct properties and combining them helps to learn useful and complementary representations of the data. Still, representing and fusing different modalities in an appropriate and efficient manner is an open research question.

The extraction of visual cues for emotion recognition has been receiving a great deal of attention in the past decade. Recently, with the rapid growth of Convolutional Neural Networks (CNNs), extracting visual features from video frames has been investigated in many emotion recognition tasks and there are various face pretrained models made available [34, 36, 40]. However, those models are not directly suitable for video due to the lack of the temporal information and to the variation of emotion expression patterns across individuals. To deal with this issue, 3D versions of CNN have been recently proposed [43].

Adding the audio information surely plays an important role in emotion recognition in video. Most of the multimodal approaches mainly used hand crafted audio features such as the Mel Frequency

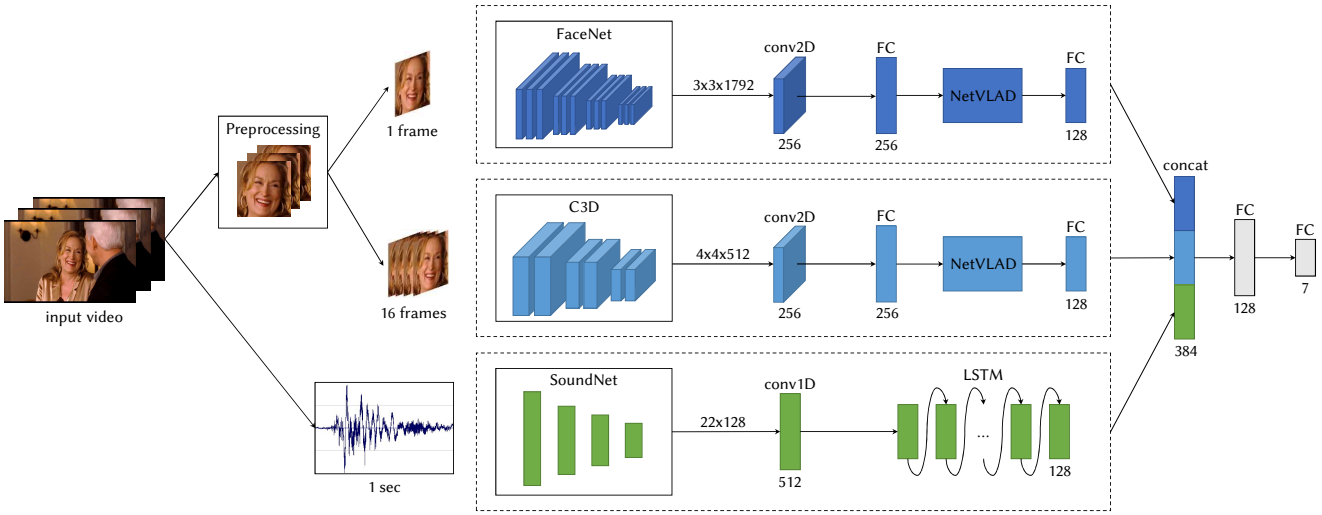


Figure 1: Overview of the proposed architecture.

Cepstrum Coefficients (MFCC) or spectrograms, with either traditional [33, 42] or deep [48] classifiers. However, those audio features are very low level and are not designed for video analysis.

In this paper, we propose a deep multimodal architecture for emotion recognition. Visual and temporal information are presented using a hybrid 2D-3D CNN approach, whereas the audio information is extracted using a deep CNN that has been trained by transferring knowledge from vision to sound [2]. To the best of our knowledge, learned deep audio features have not been yet investigated in the context of multimodal emotion recognition. The remainder of the paper is organized as follows: Section 2 presents related work, Section 3 describes the proposed multimodal emotion recognition architecture, Section 4 presents experiments and results, and finally Section 5 concludes the work and gives some future directions.

2 RELATED WORK

Emotions are displayed in video by visual and vocal means. Visual information is related to the dynamics patterns of face while the vocal information relies on audio signals. Recently, several deep audio-visual emotion recognition approaches have been proposed.

In this section, we briefly review the related work regarding the emotion recognition in videos, embracing the deep learning representations of appearance, temporal and audio information and the related multimodal fusion schemes.

Spatio-temporal evolution of facial features is one of the strongest cues for emotion recognition. Prior works using Deep Neural Networks (DNNs) for emotion recognition in video have mainly relied on temporal averaging and pooling strategies [5, 24]. More recently, we note an increase in using temporal neural networks such as Recurrent Neural Networks (RNN) to quantify the visual motion. Several previous works trained temporal neural network models on visual hand-crafted features [17, 35]. Few works have considered combining CNNs with RNNs [11, 26]. For instance, in the work of [11], the authors combine RNN with CNN to model the facial expression dynamic in video. The later suggested that temporal information integration improves classification results. In similar works [6, 17, 29], the authors use Long Short-Term Memory (LSTM) cells to aggregate CNN features over time. Other recent works model the motion information using 3D convolutional networks (C3D) [12, 48].

Regarding the audio information, deep learning-based approaches have recently attracted increasing attention among the computer vision community. Classical approaches rely on extracting audio hand-crafted features and apply a DNN classifier on those features. For instance, [29] and [14] investigate the use of deep learning approaches for emotional speech recognition. [14] train a DNN with MFCC features to classify emotions into 6 classes. In [47], the authors extract Mel-spectrogram features from audio signals for each video segment to classify emotions using a DNN. In [10], the authors train an LSTM on acoustic parameter set for affective computing. In [4], the authors investigate the emotional impact of movie genre to predict media interestingness. The later work use Soundnet and VGG features for genre recognition. Few works proposed to learn deep audio model from scratch and most of them are dedicated to specific task such as speech recognition [18, 21, 30]. One of the main challenges to build deep audio models is the lack of labeled sound data. For instance, in [41], the authors present a new deep architecture with data augmentation strategy to learn a model for audio events recognition. The later claims that combining visual features with deep audio features leads to significant performance in action recognition and video highlight detection compared to either the use of visual features alone or the fusion with MFCC features.

Multimodal data fusion remains an important challenge in emotion recognition systems. Previous works in multimodal emotion recognition using deep learning assume independence of different modalities, performing either early fusion (feature-level fusion) [7] or late fusion (decision-level fusion) [10, 11, 24, 44]. Fan *et al.* [12] combine RNN and C3D network in a late-fusion fashion.

Figure 2: Some examples of cropped faces extracted from input video frames.

The CNN-RNN, C3D and audio SVM model were trained separately and their prediction scores were combined into the final score. Kaya *et al.* [25] combine audio-visual data with least squares regression based classifiers and weighted late fusion scheme. Recent work investigate the use of DNN to fuse multimodal information. One advantage of DNNs is their capability to jointly learn feature representations and appropriate classifiers [27]. Some fusion methods based on fully connected layers have been suggested to improve video classification by capturing the mutual correlation among different modalities. For example, in [47], a fusion network is trained to obtain a joint audio-visual feature representation.

3 PROPOSED METHOD

To deal with the multimodal and temporal nature of the emotion recognition task, we build a network which is able to jointly extract static and dynamic features from different modalities, and to address the temporal evolution of the video.

Our architecture, as illustrated in Figure 1, is composed of three network branches, where the first and second ones are explicitly designed to deal with the visual features from the video, while the third one processes the audio of the input video clip. In particular, the first branch is a 2D CNN that processes the single frames, the second one is a C3D network that processes short frame snippets, and the third one is a 1D CNN that processes audio snippets.

Since all the branches can process either one frame or a short sequence, a temporal fusion strategy is devised, to deal with videos of varying length and to exploit temporal dependencies. In particular, the features of the first two branches are combined in the temporal dimension using a NetVLAD layer [1], which extends the VLAD [23] aggregation technique by learning its cluster centers. In the audio branch, instead, we make use of a LSTM network to learn the temporal dependencies between consecutive audio snippets, and represent them with the last hidden state of the network. Features coming from the three branches, once aggregated over time, are finally concatenated and fed to a multimodal network which is in charge of combining the visual, the motion, and the audio information.

3.1 Data Preprocessing

Video clips from emotion recognition datasets are usually collected from classic movies and TV reality shows, so most of frames contain irrelevant or misleading information, like background objects and background motion. Therefore, it is beneficial to pre-process the original video frames in order to limit this effect. Indeed, we extract all faces from each frame of the input video clip, and retain only the face bounding boxes, discarding all the rest in a frame. For the face detection and extraction phase, we use a cascaded convolutional neural network [46] in which faces are detected by means of a multi-task convolutional network which jointly detects facial landmarks and predicts the face bounding box. If more than one face is detected, we only use the crop of the biggest one as input to the model. Some examples of the performed pre-processing on video frames are shown in Figure 2. Furthermore, we follow the work of Aytar *et al.* [2] to extract and pre-process the audio information. Hence, we sampled the audio from video clips at a frequency of 22050 Hz and we saved every clip in mp3 format, single channel. Then, the waveform of every sample is scaled to be in the range $[-256, 256]$.

3.2 Hybrid Deep Visual Features Extraction

In order to capture visual and motion features, we design two different branches: the first one, based on a recent version of the popular FaceNet architecture [36], captures a set of visual features representing the face, while the second one, built upon the C3D network [43], jointly captures visual and motion information.

CNN Branch. In this branch, the Inception-ResNet v1 [40] network, trained as proposed in [36], is used as feature extractor. The network takes a color image of size 160×160 as input. The output of the fifth `Inception-resnet-C` block (of size $3 \times 3 \times 1792$) is then used as input to a small neural network composed by a convolutional layer with 256 filters of size 3×3, a fully connected layer with 256 units and a softmax layer of 7 classes. This network is trained using every extracted face from the challenge dataset and every image of the FER-2013 dataset [13] (more details regarding the datasets are available in Section 4.1). Images are preprocessed accordingly to the chosen open-source implementation of the network[1]. Note that only the last convolutional and fully connected layers are trained from scratch.

C3D Branch. In this branch, the C3D network [43] is used as feature extractor. The network takes 16 frames of size $112 \times 112 \times 3$ as input. The output of the `Pool5` block (of size $4 \times 4 \times 512$) is given as input to a small neural network composed by a max pooling layer

[1]Face Recognition using Tensorflow: https://github.com/davidsandberg/facenet/

of size 2×2 and stride 2, a convolutional layer with 256 filters of size 2×2, a fully connected network with 256 units and a softmax layer of 7 classes. The network is trained using the challenge dataset only. Slices of 16 extracted faces are used as input to the network and only the last convolutional and fully connected layers are trained from scratch.

3.3 Deep Acoustic Features Extraction

In the audio branch, the SoundNet network [2] is used as feature extractor. The output of the conv4 block (of size 22×128) is given as input to a network defined as follows. The first layer is a 1D convolutional layer with 512 filters of size 4, applied with a stride of 4. The size of the output is 6×512. The six feature vectors of size 512 can be seen as the compression of the audio temporal input, therefore they still contain temporal information. Based on that, the six feature vectors are given as input of an LSTM [20] layer with two levels and 128 hidden units. This layer is followed by a softmax layer of 7 classes. The network is trained using the challenge dataset and part of the eNTERFACE dataset [32] (more details regarding the datasets are available in Section 4.1). Audio raw waveform sequences extracted from the videos are used as input to the network. Even in this case, only the last convolutional, LSTM, and softmax layers are trained from scratch.

3.4 Temporal Aggregation and Multimodal Fusion

The fusion network has a double purpose: to combine the temporal information of the visual and motion features and to fuse the multiple modalities. In order to combine features extracted at different timesteps, the CNN branch and the C3D branch are followed by a NetVLAD layer [1].

Given a set of D-dimensional features $\{\mathbf{x}_i\}$, the layer can learn K cluster centers $\{\mathbf{c}_i\}$ in the same space of the features, and produce an aggregated description of the set with size $K \times D$, through the sum of residuals with respect to the cluster centers. Formally, the k-th row of the aggregated description is given by

$$\phi\Big(\{\mathbf{x}_i\}_{i=1}^{N}, \{\mathbf{c}_i\}_{i=1}^{K}\Big)(k) = \sum_{i=1}^{N} \delta(\mathbf{x}_i, \mathbf{c}_k) \cdot (\mathbf{x}_i - \mathbf{c}_k) \tag{1}$$

where $\delta(\mathbf{x}_i, \mathbf{c}_k)$ denotes the degree of membership of descriptor $\mathbf{x}_i$ to cluster $\mathbf{c}_k$. The resulting matrix is then column-wise L_2-normalized, flattened and then L_2-normalized again. Since an hard assignment of features to clusters would be non-differentiable, the NetVLAD layer employs a soft-assignment variant, in which $\delta(\mathbf{x}_i, \mathbf{c}_k)$ is computed as

$$\delta(\mathbf{x}_i, \mathbf{c}_k) = \frac{e^{-\alpha \|\mathbf{x}_i - \mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha \|\mathbf{x}_i - \mathbf{c}'_k\|^2}} \tag{2}$$

where α controls the decay of the response with magnitude of the distance. In practice, the learnable cluster centers $\mathbf{c}_k$ are decoupled into two sets of convolutional parameters, so that the layer can be implemented via the composition of a convolutional layer, softmax activation and the final L_2 normalizations.

In the proposed architecture, the NetVLAD layer on top of the CNN branch takes the features extracted from 48 frames and outputs an aggregated representation composed by 8 visual feature vectors corresponding to 8 different clusters. The feature vectors are then flattened and followed by a fully connected layer with 128 units to reduce the output dimension. In the C3D branch, instead, the NetVLAD layer takes 32 motion features and outputs an aggregated representation of 8 motion feature vectors corresponding to 8 different clusters. As before, the feature vectors are flattened and followed by a fully connected layer with 128 units.

Regarding the audio information, we take only one second in the middle of the video, since we found that this amount of data is sufficient to perform a good classification without over-fitting the training data. Then, the output of the two fully connected layers of the CNN and C3D branch and the output of the LSTM layer of the audio branch are concatenated forming a 384 feature vector. The obtained feature vector is followed by a fully connected layer with 128 units and a softmax layer with 7 classes. The highest output of the softmax layer is our classification of the video.

3.5 Training

The training process is composed by two phases. During the first phase, the last layers of the three branches of our architecture are trained separately. Then, the multimodal fusion network is trained using the trained branches, without the softmax layer, as feature extractors. This approach allows us to use additional datasets during the training of the single branches, obtaining more robust and generalizing networks as feature extractors.

The categorical cross-entropy loss function on the seven classes of the challenge dataset is used to train all the networks of the architecture except the multimodal fusion of the Submission 4. The fusion network related to the last submission is trained using a weighted version of the categorical cross-entropy loss function. In order to increase the importance of the most frequent classes and reduce the importance of the less frequent ones, the standard loss value is multiplied by a regularizing parameter based on the distribution of the seven classes in the training set. Specifically, an exponential function is sampled following the classes distribution on the training set to obtain the regularizing parameters.

The standard and the weighted loss function are as follows:

$$L = -\sum_{i=1}^{N} \mathbf{t}_i^T \log(\mathbf{p}_i) \tag{3}$$

$$L' = -\sum_{i=1}^{N} \lambda_{c_i} \mathbf{t}_i^T \log(\mathbf{p}_i) \tag{4}$$

where N is the number of examples in the batch, $\mathbf{t}_i$ is the target probability vector of sample i (*i.e.* a one-hot vector), $\mathbf{p}_i$ is the vector containing the predicted probabilities for sample i, c_i is the ground truth class of sample i, and λ_k is the regularizing parameter for the class k.

4 EXPERIMENTS AND RESULTS

In this section, we firstly describe the datasets used during the experiments. Then, we detail the implementation of the proposed model. Finally, we report and discuss the results achieved on the validation and the testing data.

4.1 Datasets

We trained our networks with different emotion datasets, evaluating them on the challenge dataset only.

Acted Facial Expressions in the Wild (AFEW). The Acted Facial Expressions in the Wild (AFEW) dataset [9] (2017 edition) is the dataset of the Emotion Recognition in the Wild (EmotiW) challenge. It is composed by 1809 video clips extracted from movies and, since 2016, TV series. Every clip is annotated with one of seven emotions (*Angry*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise*, and *Neutral*), but only the annotations of the training and validation sets are publicly available. Some statistics about the dataset are available in Table 1.

Table 1: Statistics of the AFEW dataset [9].

Emotion	Train	Validation	Test
Angry	133	64	98
Disgust	74	40	40
Fear	81	46	70
Happy	150	63	144
Neutral	144	63	193
Sad	117	61	80
Surprise	74	46	28
Total	773	383	653

In Figure 2, some frames from the dataset and the corresponding cropped faces are shown. The dataset is used for both training and evaluation of all the branches and the multimodal fusion.

Facial Expression Recognition 2013 (FER-2013). The Facial Expression Recognition 2013 (FER-2013) dataset [13] has been created for the Facial Expression Recognition Challenge. 35,887 grayscale images have been crawled on the web and annotated with one of seven emotions (*Angry*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise*, and *Neutral*). These additional images increase the accuracy of the CNN branch when used during the training.

eNTERFACE. The eNTERFACE dataset [32] consists of 1166 video clips annotated with one of the six basic emotions (*Angry*, *Disgust*, *Fear*, *Happy*, *Sad* and *Surprise*). The clips are recorded in constrained environments and contain both audio and video data. This dataset is used during the training of the Audio branch of our architecture, decreasing the over-fitting.

4.2 Implementation Details

Detected faces are pre-processed and resized to comply with the expected input of the CNN and C3D network, respectively 160×160 and 112×112. Data augmentation techniques are applied during the training in order to reduce the over-fitting and increase the generalization capabilities of our architecture. Random flip, crop, and zoom are applied to the visual input, while the audio is used "as it is", but a random 1 second-length slice is selected every time.

Furthermore, batch normalization [22] is applied before every activation of the trained layers in conjunction with dropout [19, 39] between fully connected layers. Dropout is also applied on the LSTM block of the Audio branch, following [45] and [37]. The related keep probabilities range between 0.1 and 0.8 based on the position and the network where dropout is applied. Low keep probability values allow to reduce the over-fitting despite the small amount of training data.

The parameters of the layers for which we did not use pre-trained weights are initialized following what was proposed in [16]. The Adam optimizer [28] is used during the training in every network of our architecture with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate varies depending on the network, due to the considerable differences between the branches and the fusion network. Regarding the CNN, C3D, Audio, and fusion network, the learning rate are respectively 0.0001, 0.001, 0.001, and 0.0005. In all the networks, the target batch size is 128, but we are forced to drastically reduce it in the C3D and fusion network due to memory space limits.

4.3 Results

In this section, we present the achieved results using our different approaches on the challenge validation and test sets.

4.3.1 Results on Validation Set. To validate the performance of our models, we conduct first a set of experiments on the validation set. Table 2 presents the best achieved results for the single branches and the multimodal fusion approaches. For the CNN and C3D branch, the accuracy is reported with respect to both every single frame and the whole video. In the latter case, the prediction of every video is obtained averaging the predictions of its frames.

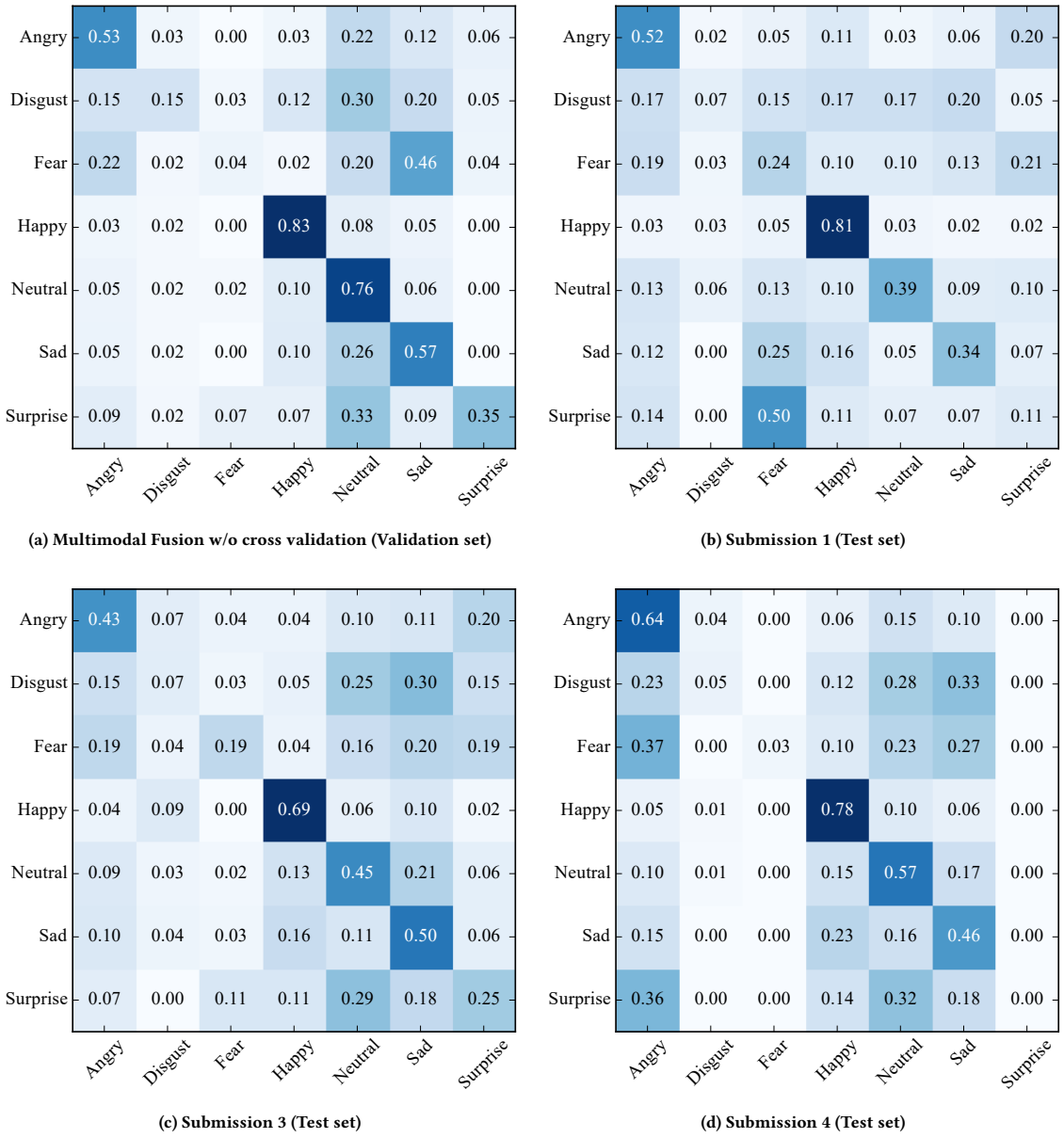
Table 2: Experimental results on the AFEW validation set.

Model	Accuracy (%)
CNN Branch (single frame)	39.95
CNN Branch (average)	**44.50**
C3D Branch (single slice)	33.31
C3D Branch (average)	31.59
Audio Branch (1 second)	33.65
Multimodal Fusion (whole video)	**50.39**

As one can note, the multimodal fusion gives an absolute accuracy gain of nearly 6% with respect to the best single branch. Both the temporal combination on the visual branches and the multimodal fusion of the three branches contribute to the accuracy improvement. The corresponding confusion matrix is shown in Figure 3a.

It is worth to notice that the proposed architecture is able to classify almost every emotion with a good accuracy on the validation set. The only exception is the class *Fear*, mainly confused with *Angry*, *Neutral*, and *Sad*. Interestingly, while analyzing results on the validation data we observed that the CNN branch correctly classify about every emotion with an acceptable accuracy, the C3D branch is unable to classify the classes *Disgusted*, *Fear*, and *Surprise* in most of the cases whereas the Audio branch never correctly classify the classes *Disgusted* and *Surprise*.

Additional experiments were made on the validation set to investigate the fine-tuning of the pre-trained networks (FaceNet, C3D,



(a) Multimodal Fusion w/o cross validation (Validation set)

(b) Submission 1 (Test set)

(c) Submission 3 (Test set)

(d) Submission 4 (Test set)

Figure 3: Confusion matrices on the AFEW validation and test sets [9].

SoundNet), but they resulted in an early over-fitting of the models. Indeed, over-fitting has been a major issue in most of the experiments performed in this work. We think this is mainly caused by the limited size of the available datasets regarding multimodal emotion recognition. Furthermore, most of the available datasets contain video clips recorded in constrained environments in which a subject acts an emotion. As a result, expressed emotions are not natural and audio information is rarely available.

4.3.2 Challenge Submissions: Results on Test Set. In order to evaluate the performance of our different approaches/models on the challenge test set, we submitted 4 runs to EmotiW 2017 challenge. In this paper, we present only the three best submissions (1, 3 and 4). In particular, submission 1 corresponds to a preliminary version of our architecture: it contains the CNN branch and the Audio branch only and it is trained with the standard categorical cross-entropy loss function (Eq. (3)). Differently, both submission 3 and submission 4 correspond to the architecture described in Section 3. The standard categorical cross-entropy loss function (Eq. (3)) is used in the first case while its weighted version (Eq. (4)) is used in the second one.

To increase the training data while keeping a stopping condition, the validation set of the AFEW dataset was split in five folds and the models are trained five times, following the k-fold cross validation technique. The folds were created maintaining the train and the validation fold subject-independent. Submission 2 attempted to keep emotion-balance instead of subject independence while training the same architecture as submission 1, but poor results were obtained and hence are not reported.

Table 3: Experimental results of our three best submissions.

Submission	Cross Validation (%)	Test Set (%)
1	49.61	44.87
3	**53.49**	44.56
4	48.30	**49.92**

The results of our submissions are presented in Table 3. The second column of the table contains the averaged accuracy on the five validation folds, while the third one contains the results of our submissions on the test set.

Looking at Figure 3 and Table 3, it can be noticed that the multimodal fusion network trained with the weighted loss (Submission 4) performs better on the test set, while the model trained with the standard loss performs better on the five validation folds. This counter-intuitive behaviour is presumably attributable to the different class distribution of the test set of the AFEW dataset compared to the train and validation set of the same dataset. Moreover, the test set contains video clips extracted from TV series (since 2016), while the training and the validation set don't. We think that these are the reasons of the discrepancy between our results on the validation set and on the test set.

Table 4: Proposed method accuracy compared to the challenge baseline.

Method	Validation Set (%)	Test Set (%)
Challenge baseline	38.87	40.47
Proposed method	**50.39**	**49.92**

As shown in Table 4, our deep learning-based architecture outperforms by a clear margin the challenge baseline [8] both on validation and test set. In particular, our best submission reaches an accuracy of 49.92%, corresponding to an absolute improvement of 9.45% with respect to the challenge baseline.

5 CONCLUSION

We proposed a multimodal deep learning framework for emotion recognition in video that participated to the audio-video based subchallenge of the Emotion Recognition in the Wild 2017 challenge. Our approach combines visual, temporal and audio information using neural network-based architectures only. Notwithstanding the small amount of labelled data regarding the emotion recognition in video, the proposed method outperforms the challenge baselines of 38.87% and 40.47% obtaining an accuracy of 50.39% and 49.92% on the validation and the test dataset respectively. In the future, making use of larger annotated datasets, we are planning to train the entire architecture in one step and to fine-tune the pre-trained audio and visual models to make the extracted features more domain specific.

ACKNOWLEDGMENTS

The research leading to this paper was partially supported by Bpifrance within the NexGen-TV Project, under grant number F1504054U.
We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

REFERENCES

[1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*.
[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Neural Information Processing Systems*.
[3] Emile Barkhof, Leo M.J. de Sonneville, Carin J. Meijer, and Lieuwe de Haan. 2015. Specificity of facial emotion recognition impairments in patients with multi-episode schizophrenia. *Schizophrenia Research: Cognition* (2015).
[4] Olfa Ben-Ahmed, Jonas Wacker, Alessandro Gaballo, and Benoit Huet. 2017. EURECOM @MediaEval 2017: Media Genre Inference for Predicting Media Interestingnes. In *the Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017.*
[5] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2014. Multi-scale Temporal Modeling for Dimensional Emotion Recognition in Video. In *International Workshop on Audio/Visual Emotion Challenge.*
[6] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2016. Long short term memory recurrent neural network based encoding method for emotion recognition in video. In *IEEE International Conference on Acoustics, Speech and Signal Processing.*
[7] Shizhe Chen and Qin Jin. 2015. Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks. In *International Workshop on Audio/Visual Emotion Challenge.*
[8] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. In *ACM International Conference on Multimodal Interaction.*
[9] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* (2012).
[10] Wan Ding, Mingyu Xu, Dongyan Huang, Weisi Lin, Minghui Dong, Xinguo Yu, and Haizhou Li. 2016. Audio and Face Video Emotion Recognition in the Wild Using Deep Neural Networks and Small Datasets. In *ACM International Conference on Multimodal Interaction.*
[11] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent Neural Networks for Emotion Recognition in Video. In *ACM International Conference on Multimodal Interaction.*
[12] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In *ACM International Conference on Multimodal Interaction.*
[13] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. 2013. Challenges in Representation Learning: A report on three machine learning contests. In *International Conference on Machine Learning Workshops.*
[14] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In *Fifteenth Annual Conference of the International Speech Communication Association.*
[15] A. Hargreaves, O. Mothersill, M. Anderson, S. Lawless, A. Corvin, and G. Donohoe. 2016. Detecting facial emotion recognition deficits in schizophrenia using dynamic stimuli of varying intensities. *Neuroscience letters* (2016).
[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision.*
[17] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. 2015. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *International Workshop on Audio/Visual Emotion Challenge.*
[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep Neural

Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (2012), 82–97.

[19] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 801–804.

[22] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*.

[23] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

[24] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, and Yoshua Bengio. 2013. Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video. In *ACM International Conference on Multimodal Interaction*.

[25] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* (2017).

[26] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S Huang. 2016. How Deep Neural Networks Can Improve Emotion Recognition on Video Data. In *IEEE International Conference on Image Processing*.

[27] Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[28] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* (2014).

[29] Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and Recurrent Neural Networks. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

[30] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. 2014. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Trans. Multimedia* 16 (2014), 2203–2213.

[31] Francesco Marchi and Albert Newen. 2015. Cognitive penetrability and emotion recognition in human facial expressions. *Frontiers in psychology* (2015).

[32] O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *International Conference on Data Engineering Workshops*.

[33] Marco Paleari, Benoit Huet, and Ryad Chellali. 2010. Towards Multimodal Emotion Recognition: A New Approach. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '10)*. ACM, New York, NY, USA, 174–181. https://doi.org/10.1145/1816041.1816069

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[35] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *International Workshop on Audio/Visual Emotion Challenge*.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

[37] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2016. Recurrent Dropout without Memory Loss. *CoRR* abs/1603.05118 (2016).

[38] John R. Smith, Dhiraj Joshi, Benoit Huet, Hsu Winston, and Jozef Cota. 2017. Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation. In *Proceedings of ACM Multimedia*. October 23-27, 2017, Mountain View, CA, USA.

[39] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15, 1 (2014), 1929–1958.

[40] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *International Conference on Learning Representations Workshops*.

[41] Naoya Takahashi, Michael Gygli, and Luc Van Gool. 2017. AENet: Learning Deep Audio Features for Video Analysis. *CoRR* abs/1701.00599 (2017).

[42] Shriman Narayan Tiwari, Ngoc QK Duong, Frédéric Lefebvre, Claire-Helène Demarty, Benoit Huet, and Louis Chevallier. 2016. Deep Features for Multimodal Emotion Classification. (2016).

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*.

[44] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen. 2016. HoloNet: Towards Robust Emotion Recognition in the Wild. In *ACM International Conference on Multimodal Interaction*.

[45] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent Neural Network Regularization. *CoRR* abs/1409.2329 (2014).

[46] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[47] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2016. Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition. In *ACM International Conference on Multimedia Retrieval*.

[48] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).